

Image Super-Resolution via Sparse Bayesian Modeling of Natural Images

Haichao Zhang, David Wipf and Yanning Zhang

Abstract

Image super-resolution (SR) is one of the long-standing and active topics in image processing community. A large body of works for image super resolution formulate the problem with Bayesian modeling techniques and then obtain its Maximum-A-Posteriori (MAP) solution, which actually boils down to a regularized regression task over separable regularization term. Although straightforward, this approach cannot exploit the full potential offered by the probabilistic modeling, as only the posterior mode is sought. Also, the separable property of the regularization term can not capture any correlations between the sparse coefficients, which sacrifices much on its modeling accuracy. We propose a Bayesian image SR algorithm via sparse modeling of natural images. The sparsity property of the latent high resolution image is exploited by introducing latent variables into the high-order Markov Random Field (MRF) which capture the content adaptive variance by pixel-wise adaptation. The high-resolution image is estimated via Empirical Bayesian estimation scheme, which is substantially faster than our previous approach based on Markov Chain Monte Carlo sampling [1]. It is shown that the actual cost function for the proposed approach actually incorporates a non-factorial regularization term over the sparse coefficients. Experimental results indicate that the proposed method can generate competitive or better results than *state-of-the-art* SR algorithms.

Index Terms

Super-resolution, natural image statistics, Markov random field, Field-of-Experts, Empirical Bayesian, non-factorial prior, non-stationary prior, Bayesian minimum mean square error estimation

I. INTRODUCTION

Super Resolution (SR) is one of the very active research topics in image processing community. In practice, many tasks such as astronomical observation and medical diagnostic rely on high quality images for reliable and accurate analysis as well as prediction. However, in many practical situations, due to the inherent limitation of the optical system or other factors, the observed images are often of low resolution,

thus limiting the subsequent tasks based on them. Image SR aims to estimate a high-resolution image (HR) from a single or a set of low-resolution (LR) observations [2], which is termed as single image SR and multi-frame SR respectively. SR has a long history and is a still very active field [2], with many new techniques emerging [3], [4], [5], [6].

Conventional multi-frame SR methods are reconstruction based methods. They typically perform motion estimation followed by frame fusion [2], which rely on the complementary information contained in different frames for resolution enhancement. The success of this kind of approaches relies heavily on the accuracy of the motion estimation procedure as well as the later frame fusion scheme. Furthermore, such reconstruction based methods are inherently limited to the case of small zooming factors [7]. For single image SR, the simplest methods are the interpolation-based approaches, *e.g.*, bicubic interpolation. However, such analytical interpolation based methods do not exploit the underlying structures in natural images, such as edges, thus usually blurring the fine details and introducing artifacts in the interpolated results. More advanced interpolation-based methods take the underlying structures of the image into consideration during interpolation, *e.g.*, [8], [9], thus improving the interpolation quality by adjusting the interpolation scheme according to the latent structures. Recently, methods exploiting the natural image priors for SR have been proposed in the literature [3], [6], [10], [11], [1]. Freeman *et al.* exploited the patch similarity prior with the help of a large training image set for SR [3]. Tappen *et al.* developed a Markov Random Field based SR algorithm utilizing the sparse derivative prior of natural images [10]. Yang *et al.* proposed a SR approach based on the sparse representation prior of image patches with respect to a properly chosen dictionary [6]. Kim *et al.* developed a regression-based SR algorithm based on kernel regression and with a sparsity prior of natural images [11]. Apart from the sparsity prior, SR methods exploiting other properties of natural images such as self-similarities and local/non-local regularities have also been proposed in the literature [4], [5]. Another property of natural images is the edge statistics, which have been exploited in [12], [13] and [14] as a prior for SR. While exploiting the natural image statistical properties as a prior, these methods typically formulate the problem within the Bayesian framework and then estimate the final HR image with a Maximum A Posteriori (MAP) estimation approach.¹ Although it is demonstrated effective in some applications, this kind of approach discards the further potential enabled by the Bayesian probabilistic modeling. Another limitation is that the regularization term is typically designed empirically *a priori*, thus is not guaranteed to be compatible

¹Some of the approaches may be formulated as a regularized regression problem, the solution to which is related to the MAP solution to the corresponding Bayesian formulation. See Section II for more details.

with the image statistics. There are indeed some Bayesian SR approaches in the literature that do not use MAP solution [15], [16], [17], [18]; however, they typically use very simple image priors to ensure the tractability of computation, thus limiting their performance. A SR approach using posterior based sampling is proposed in [15]. This method is limited by the fact that it uses a very simple Ising model as its prior [15]. Multi-frame SR methods based on Variational Bayes (VB) technique have been developed in the past [16], [17], [18]. In [16], the authors used the Total Variation (TV) model as the prior, which prefers piecewise linear images thus is not quite adequate for general natural images. In [17], the authors used a casual Gaussian MRF model as the prior, which includes latent variables indicating the dependency of two adjacent pixels. Although the model has taken the edge structure of natural images into consideration, it is not flexible enough to capture the statistics of natural images well. In [1], a Bayesian image SR method with high-order MRF modelling of natural images is proposed and the HR image is estimated via MCMC sampling. To sum up, the conventional SR approaches using regularized regression are limited by the fact that they do not exploit the full potential offered by the probabilistic modeling, although various kinds of regularization terms have been designed; on the other hand, for the previously developed Bayesian SR approach, the prior is typically limited to very simple models to ensure the tractability in computation, thus limiting their estimation quality. Furthermore, VB technique is typically used for the posterior mean estimation in those approaches, which actually compromises the modeling accuracy, while the sampling based approach is time consuming with high computational cost.

We propose in this paper a single image super-resolution algorithm using sparse bayesian modeling of natural images, which enjoys several advantages:

- 1) the proposed method is derived in a Bayesian framework, which can incorporate the *a priori* knowledge on the latent HR image as well as other uncertainties such as the noise level into the framework in a natural way;
- 2) a high-order MRF model is used for capturing the HR image statistics. Different from [1] which uses a Gaussian Scale Mixtures (GSM) with finite discrete scales, our model uses an infinite number of components in the GSM model, with the scales adaptively learned along with the process of HR image estimation, rather than chosen from a pre-specified finite set of discrete scales as in [1]. Also, as a global statistical prior is used for the HR image, our algorithm does not require to re-train the model when the zooming factor is changed as required by the sparse representation based SR method [6]. Therefore, the proposed method is more flexible.
- 3) an empirical Bayesian method is applied for the task of HR image estimation as well as the

estimation of the hyper-parameters, which does not require ad-hoc modifications (tuning the trade-off parameters) as the MAP approach to achieve desirable restoration performance. Also, it is less sensitive to the local minima in the solution space than MAP, especially when heavy-tailed priors are applied. The proposed method is much more efficient than the method based on Markov Chain Monte Carlo (MCMC) sampling [1] and can generate results comparable to or better than that method.

The rest of this paper is organized as follows. We first give a review of some related previous works briefly in Section II, including the general restoration framework as well as the previous approaches for modeling natural images. We introduce our SR framework with Sparse Bayesian modeling of natural images in Section III. Then we present an efficient and effective super-resolution algorithm in Section IV. Experiments are conducted in Section V and the results are compared with several *state-of-the-art* algorithms. We conclude this paper in Section VI.

II. PREVIOUS WORKS

We review some of the related previous works in this section, which will lay the foundation for the derivation of our approach later.

A. Image Restoration Model

Mathematically, the LR image formation process is usually modeled as the convolution of the latent HR image with a low-pass filter which models the low passing property of the imaging system followed with downsampling,

$$\mathbf{y} = \mathbf{D}\mathbf{H}\mathbf{x} + \boldsymbol{\epsilon}, \quad (1)$$

where $\mathbf{x} \in \mathbb{R}^m$ and $\mathbf{y} \in \mathbb{R}^n$ is the vector representation of the HR and LR image, respectively, obtained by lexicographical ordering the images. $\mathbf{H} \in \mathbb{R}^{m \times m}$ is the matrix corresponding to the blurring process with blur kernel \mathbf{h} and $\mathbf{D} \in \mathbb{R}^{n \times m}$ ($n < m$) is the matrix representing the downsampling operator. $\boldsymbol{\epsilon} \in \mathbb{R}^n$ is a noise term. The task of SR is to estimate the underlying HR image \mathbf{x} given only the LR observation \mathbf{y} . In the sequel, we review briefly some related works in the literature.

As the task of SR is to estimate a high resolution image from the low resolution observation, the number of unknowns m is larger than that of the knowns n , thus the problem is ill-posed. Therefore, regularization techniques based on the *a priori* knowledge are required to alleviate the ill-posedness of this problem. Typically, the task of SR is formulated as a regularized least square regression problem.

Mathematically,

$$\hat{x} = \arg \min_x \|\mathbf{y} - \mathbf{D}\mathbf{H}\mathbf{x}\|_2^2 + \lambda R(\mathbf{x}), \quad (2)$$

where the first term is the fidelity term reflecting the fact that the unknown HR image \mathbf{x} should generate observation that is similar to \mathbf{y} after passing it through the observation system. The second term, $R(\cdot)$, is a regularization term on the desired HR image to alleviate the ill-posedness of the SR problem. λ is a parameter balancing the contributions of these two terms. This is a very popular model and has been adopted in many recent SR and other related works [13], [14], [19], [20]. Following this direction, most of the current works focus on designing different formulations for the regularization term $R(\cdot)$. The most classical choice for $R(\cdot)$ would be the squared ℓ_2 -norm, which is based on the assumption that the energy of the HR image should not be too large. Because of its well known blurring effect for the ℓ_2 -norm based regularization, some algorithms apply the ℓ_2 -norm in the gradient domain. Recently, many approaches exploiting the sparsity property of natural images are developed [6], [10], [11], [13], [14], [20]. A general approach for designing $R(\cdot)$ utilizing the sparsity property is to use the sparsity property in the transform domain (*e.g.*, wavelet domain, edge domain) as regularization [13], [14]. A novel regularization term called ‘Softcuts’ has been developed in [13], by exploiting the sparse property of the edge profiles. In [14], the authors use the sparsity prior of the edges in images by first estimating the gradient of the HR image via transforming that of the LR image with the learned relationship between the LR and HR gradients. Then the regularization term $R(\cdot)$ is designed as the distance between the gradients of the unknown latent HR image and the estimated HR gradients. A sparse representation (coding) based SR (ScSR) method has been proposed in [6] based on the property that natural image patches have a sparse representation with respect to a properly chosen dictionary. Specifically, the ScSR method first recovers the sparse representation coefficients from the low resolution patches with respect to a low resolution dictionary, and then reconstructs a high resolution patch with the recovered representation coefficients and a high resolution dictionary, which is trained jointly with the low resolution dictionary. To further enhance the texture details in the estimated HR images, several approaches have been proposed by using texture prior offered by domain specific examples [20], [21], [22].

The regularized SR model (2) is related to a probabilistic model as follows:

$$\begin{aligned} p(\mathbf{x}|\mathbf{y}) &= \frac{p(\mathbf{y}|\mathbf{x})p(\mathbf{x})}{p(\mathbf{y})} \propto p(\mathbf{y}|\mathbf{x})p(\mathbf{x}) \\ &\propto \exp\left(-\frac{\|\mathbf{y} - \mathbf{D}\mathbf{H}\mathbf{x}\|_2^2}{\sigma^2}\right) \exp\left(-\frac{R(\mathbf{x})}{\eta^2}\right), \end{aligned} \quad (3)$$

where σ denotes the standard deviation of the noise and η is a scale parameter for the prior distribution

of \mathbf{x} . At this point, it is easy to see that the solution to (2) corresponds to the Maximum-A-Posteriori (MAP) solution to (3) due to the following relation

$$\hat{\mathbf{x}} = \arg \max_{\mathbf{x}} p(\mathbf{x}|\mathbf{y}) \Leftrightarrow \arg \min_{\mathbf{x}} -\log p(\mathbf{x}|\mathbf{y}), \quad (4)$$

which corresponds to the mode of the posterior distribution for the HR image \mathbf{x} . The regularization parameter λ in (2) is related to σ and η in (3) by $\lambda = \frac{\sigma^2}{\eta^2}$.

Because of this relation, many previous works first model the SR problem in a probabilistic framework as (3) and then estimate the HR image by minimizing the negative logarithm of it, which actually boils down to the regularized regression model as shown in (2) [14], [20]. To solve (2), gradient based optimization schemes are typically used [14], [20]. While some encouraging results can be obtained, the MAP approach could not exploit the full potential offered by the probabilistic modeling, as only the posterior mode is sought for solution, discarding much information offered by the probabilistic model. Even worse, it is prone to getting stuck in a local minima when the prior model is a heavy-tailed distribution, which actually leads to a non-convex optimization problem, thus there is no guarantee of obtaining even the posterior mode.

B. Prior Models based on Natural Image Statistics

The statistics of natural images have received great attention of researchers from different communities for both understanding the human visual system and designing more effective processing algorithms [23]. Natural image statistics modeling has a fundamental relation with the field of image processing. Natural images are different from random noise image in that they exhibit structures. Examples are local regularities such as edges [9] and self-similarities [4], [5]. As a result, the natural images only occupy a tiny fraction of the whole space of all the images generated by all the possibilities enables by the pixel value combinations. Natural images have strong statistical properties known as natural image statistics [24], [25], [26]. One of the most well-known properties is that the natural images exhibit heavy-tailed distribution when applying derivative filters onto them. This can be explained intuitively as follows: on one hand, natural images are locally smooth, therefore, the local difference will be small, thus the marginal distribution will decrease faster than the Gaussian distribution; on the other hand, natural images have many structures such as edges, where the derivative response can be large, which contributes to the heavier tails than the Gaussian distribution. This prior enjoys wide range of applications, including image denoising [25], [26], deblurring [27] and super-resolution [1], [11]. Apart from this heavy-tailed marginal statistics, natural images also exhibit several other properties, such as scale invariance, which

states that the natural images exhibit similar heavy-tailed distribution at different scales [24]. Another property is the joint statistics, meaning that the neighboring pixels in the natural images exhibits high dependency [26].

There are many works on image priors and it is still a very active research topic. Gaussian model applied onto the derivatives of images is the most classical and widely used one due to its simplicity:

$$p(\mathbf{x}) \propto \exp\left(-\frac{\|\nabla \mathbf{x}\|_2^2}{\eta^2}\right), \quad (5)$$

where $\nabla \mathbf{x}$ denotes the gradients of image \mathbf{x} . Gaussian prior is related to the Tikhonov regularization in the inverse problem theory. Although it has the advantage of closed-form solution, it can not generate satisfying solutions in most of the image restoration cases as it typically smooths the image too much. To preserve the edge structure, Laplacian prior is later used as image prior and is defined as:

$$p(\mathbf{x}) \propto \exp\left(-\frac{\|\nabla \mathbf{x}\|_1}{\eta^2}\right). \quad (6)$$

Laplacian prior is related to the Total Variation (TV) model derived in the Partial Differential Equation field [28], which has been proved to keep the image discontinuities better. It is also related to the ℓ_1 -norm based regularization, which is well known for its ability in promoting sparsity for the solution. Although the TV prior can preserve the edges, it can not capture natural image statistics well enough, as the resulting images are typically piecewise linear, which is desirable for cartoon-like images, but not for natural images. Natural images follow a distribution with heavier tails than Gaussian and Laplacian distribution, which has been realized and used by many researchers recently [27], [29]. To overcome this limitation, some researchers proposed to use the hyper-Laplacian sparse prior for natural images which has been successfully applied in many fields [27], [29]:

$$p(\mathbf{x}) \propto \exp\left(-\frac{\|\nabla \mathbf{x}\|_\alpha}{\eta^2}\right), \quad (7)$$

where $\|\mathbf{x}\|_\alpha$ is defined as $\|\mathbf{x}\|_\alpha = \sum_i |\mathbf{x}(i)|^\alpha$. α here is the parameter controlling the sparseness of the desired natural images. For natural images, the responses with respect to the learned sparse filters are sparse, *i.e.*, with a large number of zero elements and a few responses with large absolute values; therefore, α is typically chosen as $0.5 \leq \alpha \leq 0.8$ in the literature [29].

An elegant formulation for unifying the above mentioned probabilistic models of natural images is Markov Random Field (MRF) [26]. MRF is a type of undirected graph, where the node and edge of this graph is defined differently according to different MRF configurations. In MRF, the probability of a whole image \mathbf{x} is defined as follows based on the computation of local cliques:

$$p(\mathbf{x}) = \frac{1}{Z} \prod_{c \in \zeta} f(\mathbf{x}_{[c]}), \quad (8)$$

where ζ denotes the set of all pixel locations of the image and c refers to a particular location. $\mathbf{x}_{[c]}$ is the vector made up by the local neighborhood at c for the image \mathbf{x} , which is also referred to as a clique. $f(\cdot)$ is a potential function, which takes a clique as input and outputs the potential or energy for that clique. Z is the partition function to ensure $p(\mathbf{x})$ to be integrated to one. One typical MRF model is the so called pairwise MRF. In this model, each pixel in the image is defined as a node. Each pixel is connected with its directed neighbors.

The choice of the potential function $f(\cdot)$ is crucial for the modeling quality. The conventional choice is the Gaussian function, as in (5). With the recent development in natural image statistics, it is demonstrated that the potential function defined as a heavy-tailed function over the response of derivative filters on the clique \mathbf{x}_c can better capture the image statistics [26]. Motivated by the property of heavy-tailed marginal distribution of natural images, to improve the modeling quality, researchers suggest to use heavy-tailed potential functions such as Laplacian and hyper-Laplacian as in (6) and (7) and Student-t distribution as used in [26].

In the terminology of MRF, the models (5)~(7) are pairwise MRF, as only two neighboring pixels are involved in each clique, due to the usage of the first order derivative filters. However, the modeling power of pairwise MRF is rather limited, as the pixels in natural images also exhibit long range relations; therefore, high-order MRF model is developed in some recent work [26]. One challenge in the case of high-order MRF is that the proper potential function is hard to define in the high dimensional space, as the size of a clique has been increased. The recently introduced Field-of-Experts (FoE) model formulates the potential on clique c as a product of several functions defined on different low dimensional spaces respectively [26]:

$$\begin{aligned} f(\mathbf{x}_{[c]}) &= \prod_{l=1}^L \phi((\mathbf{k}_l * \mathbf{x})_c; \Theta_l) \\ &= \prod_{l=1}^L \phi((\mathbf{K}_l \mathbf{x})_c; \Theta_l), \end{aligned} \quad (9)$$

which is known as the Product-of-Expert (PoE) model in the literature [30]. $(\mathbf{k}_l * \mathbf{x})_c$ refers to the c -th pixel in the filtered image by \mathbf{k}_l . $\mathcal{K} = \{\mathbf{k}_l\}_{l=1}^L$ is a set of derivative filters, either designed under some criteria or learned from the data [25], [26]. $\mathbf{K}_l \in \mathbb{R}^{m \times m}$ is the matrix representation corresponding to the derivative filter \mathbf{k}_l , with the following relation: $\mathbf{K}_l \mathbf{x} = \mathbf{k}_l * \mathbf{x}$. The clique $\mathbf{x}_{[c]}$ in this model is actually the local patch covered by the filter \mathbf{k}_l ($l \in \{1, \dots, L\}$) at c . The model in (9) consists of L expert functions to model the high-dimensional distribution by using the products of them. $\{\Theta_l\}_{l=1}^L$ is the parameter set associated with the PoE.

In FoE, the filters are learned from the training data [26]. The advantage of learned filters over pre-specified ones (such as $\{\mathbf{k}_1 = [-1, 1], \mathbf{k}_2 = \mathbf{k}_1^T\}$) is that the learned filters have the potential to capture the statistics of natural images better than the simple horizontal and vertical derivative filters. Note that the prior can also be made adaptive to the content of the specific image by allowing its parameters to change according to the content [31]. One challenge in learning the MRF model is that the normalization term Z is typically intractable to evaluate, which poses a great difficulty for the learning process. To handle this challenge, several schemes such as Contrastive Divergence [30] and Score Matching [32] are proposed in the literature.

III. IMAGE SUPER RESOLUTION VIA SPARSE BAYESIAN MODELING

As mentioned above, the conventional SR approaches using regularized regression can not exploit the full potential of the probabilistic model and are also often faced with non-convex optimization problem when heavy-tailed sparse priors are used, thus is prone to be trapped in a local minima; on the other hand, previously proposed Bayesian SR approach² only used very simple prior model for natural images due to the limitation in computation [16], [17], [18], or resorting to MCMC sampling when a more advanced prior is used in the Bayesian model [1]. In this section, we present an effective and efficient Bayesian SR approach which can use advanced prior model learned from natural images as [1], while reducing the computational overhead significantly. The framework of the proposed approach is illustrated in Figure 1. In the sequel, we first formulate the SR problem in the Bayesian framework and then present an effective algorithm for generating the SR image. In terms of Bayesian inference, the latent SR image is estimated via the posterior distribution $p(\mathbf{x}|\mathbf{y}; \Theta)$. Using the Bayes rule, we can get:

$$p(\mathbf{x}|\mathbf{y}; \Theta) = \frac{p(\mathbf{y}|\mathbf{x}; \Theta)p(\mathbf{x})}{p(\mathbf{y})}, \quad (10)$$

therefore, the task for the Bayes modeling includes the configuration of the likelihood $p(\mathbf{y}|\mathbf{x}; \Theta)$ and the prior $p(\mathbf{x})$, which is presented in detail in the sequel.

A. Bayesian Image Super Resolution Formulation

In this subsection, we first present the likelihood and prior model, and then derive the posterior for the latent HR image, from which the posterior mean is computed as the estimation of the HR image.

²By Bayesian SR, we refer to those approaches that use the MMSE estimation criteria, rather than those using the MAP criteria, even though they are formulated in a Bayesian framework.

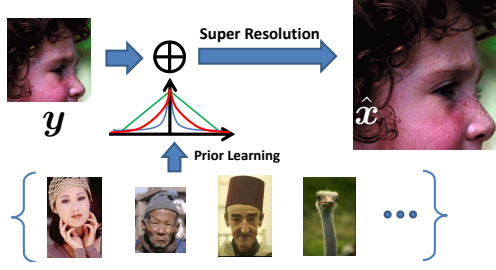


Fig. 1. **Bayesian SR Framework.** In the proposed SR framework, we first learn a set of filters from the training images with a flexible high-order MRF model, with Gaussian Scale Mixture model with continuous scale as its potentials. The learned filter is then used together with the potential model as the prior for the underlying HR image to be estimated. The proper scale is adapted to the image and is estimated together with the HR image estimation. The HR image is estimated via an empirical Bayesian approach. In this framework, we are actually ‘transfer’ the prior learned from training images to the unknown HR image.

1) *Likelihood:* The noise term ϵ is assumed to be known only up to the statistical property, *i.e.* its statistical distribution. The statistical distribution of the noise is application dependent. In many situations, the noise term can be modeled with Gaussian distribution. In this work, we also use a Gaussian distribution for modeling the statistics of noise. Assuming the elements in the noise term ϵ follow *i.i.d.* Gaussian distribution and the standard deviation is σ , then we have:

$$\epsilon \sim \mathcal{N}(\mathbf{0}, \sigma^2 \mathbf{I}). \quad (11)$$

With (11) and the observation model defined in (1), we can derive the following conditional distribution for the observed image:

$$p(\mathbf{y}|\mathbf{x}, \sigma^2) = \mathcal{N}(\mathbf{y}|\mathbf{D}\mathbf{H}\mathbf{x}, \sigma^2 \mathbf{I}), \quad (12)$$

where $\mathbf{I} \in \mathbb{R}^{n \times n}$ is the identity matrix. The noise level σ^2 is not assumed to be known. Rather, we treat it as a random variable and place a hyperprior on it as shown later.

2) *Prior Model for Image:* The definition of the prior model for the HR image is crucial for the final HR estimation quality. The design of such a prior should reflect the desired property of natural HR images. The Gaussian prior reflects the smoothness property of natural images. However, this prior typically causes blurring effects for the structures in natural images, as it fails in capturing the heavy-tailed distribution of the natural image statistics. Also, the conventional first-order MRF model for natural images are limited by the pairwise interaction, as it is not flexible enough for capturing the image statistics. The recently developed FoE model is a high-order MRF model that can capture the statistics of natural images better

by learning the high-order filters from the data [26]. Therefore, in this work, we use a high-order FoE model as the prior for the underlying HR image \mathbf{x} :

$$p(\mathbf{x}) = \frac{1}{Z} \prod_{c \in \zeta} f(\mathbf{x}_{[c]}) = \frac{1}{Z} \prod_{c \in \zeta} \left(\prod_{l=1}^L \phi((\mathbf{k}_l * \mathbf{x})_c; \Theta_l) \right), \quad (13)$$

where Gaussian scale mixtures (GSM) model (15) is used as the expert function $\phi(\cdot)$ rather than Student-t distribution as used in [26], and the Gaussian, Laplacian and hyper-Laplacian functions used in (5), (6) and (7) respectively. This model is very flexible and can be adapted easily for capturing the natural image statistics. In our previous work [1], we have used *discrete* and *finite* Gaussian Scale Mixture (GSM) as the expert function:

$$\phi((\mathbf{k}_l * \mathbf{x})_c; \Theta_l) = \sum_{j=1}^J \beta_{lj} \mathcal{N}((\mathbf{k}_l * \mathbf{x})_c; 0, \frac{\eta_l^2}{s_j}), \quad (14)$$

where β_{lj} is the normalized weight for the j -th component in the Gaussian mixture for modeling the responds of the l -th filter. η_l^2 is the base variance for the J Gaussian components in the l -th expert function. s_j is the scale parameter for the j -th component. The usage of this expert function allows efficient sampling as used in [1]. However, in this model, the scales are pre-defined and restricted to be one of the values in the discrete set.

Now we resort to variational approach for modeling the expert function $\phi(\cdot)$ as:

$$\phi((\mathbf{k}_l * \mathbf{x})_c | \gamma_{cl}) = \max_{\gamma_{cl} \geq 0} \mathcal{N}((\mathbf{k}_l * \mathbf{x})_c | 0, \gamma_{cl}) \psi(\gamma_{cl}), \quad (15)$$

which is actually the superposition of zero-mean-Gaussian with different variances.

Using (13) and (15), we can have the following prior model for the latent HR image \mathbf{x} as:

$$p(\mathbf{x} | \{\gamma_l\}_{l=1}^L, \tau) = \mathcal{N}(\mathbf{x} | \mathbf{0}, \mathbf{V}_\gamma^{-1}), \quad (16)$$

where $\mathbf{V}_\gamma^{-1} = \sum_{l=1}^L \mathbf{K}^\top \mathbf{\Gamma}_l^{-1} \mathbf{K}_l$ and $\mathbf{\Gamma}_l = \text{diag}(\gamma_l)$.

3) *Hyperpriors on Hyperparameters*: We do not assume the noise level to be known in our model. Rather, we treat the noise level as a random variable and perform estimation for it jointly with other random variables. Specifically, to model this random variable, we use a Gamma distribution as its hyperprior:

$$p(\tau) = \mathcal{G}(\tau; a^o, b^o) = \frac{1}{\Gamma(a^o)} (b^o)^{a^o} \tau^{a^o-1} e^{-b^o \tau}, \quad (17)$$

where $\tau = \sigma^{-2}$ which is also referred to as precision. a^o and b^o are non-negative parameters controlling the shape and scale of the Gamma distribution, respectively. For the latent variables $\{\gamma_l\}_{l=1}^L$, we also

place gamma priors over them, which will promote sparsity combined with the prior for \mathbf{x} :

$$p(\{\gamma_l\}_{l=1}^L) = \prod_{l=1}^L \prod_{c=1}^N \mathcal{G}(\gamma_{cl}; a, b), \quad (18)$$

where a and b are the hyper-parameters for the hyper-priors.

4) *Posterior*: By substituting (12), (16), (17) and (18) into (10), we can derive the following posterior for the latent HR image \mathbf{x} as well as the precision τ :

$$\begin{aligned} p(\mathbf{x}, \{\gamma_l\}_{l=1}^L, \tau | \mathbf{y}) \\ \propto p(\mathbf{y} | \mathbf{x}, \{\gamma_l\}_{l=1}^L, \tau) p(\mathbf{x} | \{\gamma_l\}_{l=1}^L, \tau) p(\{\gamma_l\}_{l=1}^L) p(\tau) \\ \propto \mathcal{N}(\mathbf{y} | \mathbf{D}\mathbf{H}\mathbf{x}, \tau^{-1}\mathbf{I}) \cdot \mathcal{N}(\mathbf{x} | \mathbf{0}, \mathbf{V}_\gamma^{-1}) \cdot \prod_{l=1}^L \prod_{c=1}^N \mathcal{G}(\gamma_{cl}; a, b) \cdot \mathcal{G}(\tau; a^o, b^o). \end{aligned} \quad (19)$$

IV. EMPIRICAL BAYESIAN SUPER-RESOLUTION ALGORITHM

A. Empirical Bayesian SR Algorithm

After obtaining the posterior (19), we should perform inference based on it. However, there are several challenges for this. One of the difficulties is the fact that (19) is typically intractable as we can not perform the normalization of it. Therefore, approximate schemes have to be adopted. One type of approach is to approximate the posterior distribution with separable distribution under criteria such as KL-divergence, thus transforming the problem to another easy-to-solve problem. A representative example is the Variational Bayes (VB) method [33], which has been applied to image deblurring and compressive sensing recently [27], [34]. While approaches based on the VB technique are computationally efficient as it typically offers analytic formulas for updating the variables, it compromises the modeling accuracy for computational efficiency. Apart from approximating the original model with tractable ones as in VB, another line of solution is based on generative sampling from the posterior distribution. After obtaining enough samples from the posterior distribution, the empirical mean calculated from the generated samples is used as the estimation for the posterior mean. This approach has been used in our previous work on image SR [1], which is, however, time consuming and not practical for real applications.

Rather than using VB approximation or sampling, Tipping suggested instead to decompose the posterior into two parts as follows [35]:

$$p(\mathbf{x}, \{\gamma_l\}_{l=1}^L, \tau | \mathbf{y}) = p(\mathbf{x} | \mathbf{y}, \{\gamma_l\}_{l=1}^L, \tau) p(\{\gamma_l\}_{l=1}^L, \tau | \mathbf{y}). \quad (20)$$

The reason for this decomposition is that while the full posterior is not analytically computable due to the normalization factor, the first part of (20) enables this analytical calculation. The empirical Bayesian

approach first estimates the parameters and latent variables via maximizing the marginal likelihood by integrating out the unknown variable \mathbf{x} , and then uses the estimated optimal parameters for evaluating the conditional posterior for \mathbf{x} using the first part of (20) [35], [36]. This approach is first applied in Relevance Vector Machine (RVM) as an alternative for Support Vector Machine (SVM) and also used in Automatic Relevance Determination (ARD) [35]. automatic relevance determination This approach has been introduced into single processing community for basis selection and sparse representation by Wipf *et al.* in [37], which has later been extended further in [38], [39], [36], [40], all showing superior performance in various signal processing applications. Some theoretical analysis on the optimality conditions has been carried out in [36], [40]. In this work, we further exploit the possibilities of this promising paradigm for image processing tasks. The extension, however, is non-trivial, due to the high-dimensional property inherent to image processing tasks.

Specifically, decomposing (19) into two parts according to (20), we have the first part as follows:

$$\begin{aligned} p(\mathbf{x}|\mathbf{y}, \{\gamma_l\}_{l=1}^L, \tau) &= \frac{p(\mathbf{y}|\{\gamma_l\}_{l=1}^L, \tau)p(\mathbf{x}|\{\gamma_l\}_{l=1}^L, \tau)}{p(\mathbf{y}|\{\gamma_l\}_{l=1}^L, \tau)} \\ &= \mathcal{N}(\mathbf{x}|\boldsymbol{\mu}_x, \boldsymbol{\Sigma}_x). \end{aligned} \quad (21)$$

where the posterior mean and covariance are:

$$\begin{aligned} \boldsymbol{\mu}_x &= \tau \boldsymbol{\Sigma}_x \mathbf{H}^\top \mathbf{D}^\top \mathbf{y}. \\ \boldsymbol{\Sigma}_x &= \mathbf{W}_\gamma^{-1} = \left(\tau \mathbf{H}^\top \mathbf{D}^\top \mathbf{D} \mathbf{H} + \sum_{l=1}^L \mathbf{K}_l^\top \boldsymbol{\Gamma}_l^{-1} \mathbf{K}_l \right)^{-1} \\ &= \left(\begin{bmatrix} \mathbf{H}^\top \mathbf{D}^\top & \mathbf{K}_1^\top & \dots & \mathbf{K}_L^\top \end{bmatrix} \begin{bmatrix} \tau \mathbf{I} & \dots & \mathbf{0} \\ & \boldsymbol{\Gamma}_1^{-1} & \\ \vdots & & \ddots \\ \mathbf{0} & \dots & & \boldsymbol{\Gamma}_L^{-1} \end{bmatrix} \begin{bmatrix} \mathbf{D} \mathbf{H} \\ \mathbf{K}_1 \\ \vdots \\ \mathbf{K}_L \end{bmatrix} \right)^{-1} = \left(\mathbf{K}^\top \mathbf{Z} \mathbf{K} \right)^{-1}, \end{aligned} \quad (22)$$

where $\boldsymbol{\Gamma}_l = \text{diag}(\gamma_l)$. Basically, the first part guarantees that the posterior mean estimation of the HR image can be obtained via (22) once the hyper-parameters $(\{\gamma_l\}_{l=1}^L, \tau)$ are known.

The estimation of these hyper-parameters are obtained by minimizing the negative log of the marginal likelihood with the unknown latent HR image \mathbf{x} integrated out is as follows [35], [36], [40], [41]:

$$\{\hat{\gamma}_l\}_{l=1}^L, \hat{\tau}x = \arg \min_{\{\gamma_l\}_{l=1}^L, \tau} \mathcal{L}(\{\gamma_l\}_{l=1}^L, \tau), \quad (24)$$

where $\mathcal{L}(\{\gamma_l\}_{l=1}^L, \tau)$ is the negative log of the marginal likelihood defined as

$$\begin{aligned}\mathcal{L}(\{\gamma_l\}_{l=1}^L, \tau) &= -\log \int p(\mathbf{y}|\mathbf{x}, \{\gamma_l\}_{l=1}^L, \tau) p(\mathbf{x}|\{\gamma_l\}_{l=1}^L, \tau) d\mathbf{x} \\ &= -\log p(\mathbf{y}; \{\gamma_l\}_{l=1}^L, \tau) \\ &= \mathbf{y}^\top \Sigma_{\mathbf{y}}^{-1} \mathbf{y} + \log |\Sigma_{\mathbf{y}}| + \sum_{l=1}^L f(\gamma_l) + \sum_{i=1}^n [\log \tau + f(\tau)],\end{aligned}\tag{25}$$

where $\Sigma_{\mathbf{y}} = \tau^{-1} \mathbf{I} + \mathbf{D} \mathbf{H} \mathbf{V}_{\gamma}^{-1} \mathbf{H}^\top \mathbf{D}^\top$. This leads to the following updating rules for $\{\gamma_l\}_{l=1}^L$ and τ . The optimal γ_l can be computed as [35], [36], [40]:

$$\gamma_{li} = \frac{u_{li}^2 + z_{li} + 2b}{1 + 2a},\tag{26}$$

where $\mathbf{u}_l \triangleq \mathbf{K}_l * \boldsymbol{\mu}_{\mathbf{x}}$ and the optimal z_{li} is given by:

$$\begin{aligned}z_{li} &= \frac{\partial \log |\tau \mathbf{K} \mathbf{H}^\top \mathbf{D}^\top \mathbf{D} \mathbf{H} \mathbf{K}^\top + \boldsymbol{\Gamma}^{-1}|}{\partial \gamma_{li}} \\ &= [(\tau \mathbf{K}_l \mathbf{H}^\top \mathbf{D}^\top \mathbf{D} \mathbf{H} \mathbf{K}_l^\top + \boldsymbol{\Gamma}^{-1})^{-1}]_{ii}.\end{aligned}\tag{27}$$

The optimal τ can be computed as [41]:

$$\tau = \frac{\frac{n}{2} + a^o}{\frac{1}{2} \|\mathbf{y} - \mathbf{D} \mathbf{H} \mathbf{x}\|_2^2 + 2w + b^o}.\tag{28}$$

where

$$\begin{aligned}w &= \frac{\partial \log |\tau \mathbf{H}^\top \mathbf{D}^\top \mathbf{D} \mathbf{H} + \mathbf{V}_{\gamma}^{-1}|}{\partial \tau} \\ &= \sum_{i=1}^n [\mathbf{D} \mathbf{H} (\tau \mathbf{H}^\top \mathbf{D}^\top \mathbf{D} \mathbf{H} + \mathbf{V}_{\gamma}^{-1}) \mathbf{H}^\top \mathbf{D}^\top]_{ii}^{-1}.\end{aligned}\tag{29}$$

B. Covariance Estimation

The estimation of the covariance $\boldsymbol{\Sigma}_{\mathbf{x}}$ is required for updating both γ_l and τ , as shown in (27) and (29). However, as shown in (23), explicit calculation of $\boldsymbol{\Sigma}_{\mathbf{x}}$ requires the inversion of a very large matrix, which is intractable typically and is also the case in our task. In the past, different approaches have been proposed for estimating the covariance matrix approximately, including the low-rank approximation method using Lanczos algorithm [42] and sampling based method [43], [44], which are time-consuming. In the sequel, we will present an efficient approximate covariance estimation approach. Specifically, we use a diagonal approximation of $\boldsymbol{\Sigma}_{\mathbf{x}}$ as $\tilde{\boldsymbol{\Sigma}}_{\mathbf{x}}$:

$$\tilde{\boldsymbol{\Sigma}}_{\mathbf{x}} = \text{diag} \{ \text{diag}(\boldsymbol{\Sigma}_{\mathbf{x}}) \}.\tag{30}$$

Therefore, the key is to compute $\text{diag}(\Sigma_x)$. To achieve this goal, we take advantage of the following two relations:

$$\text{diag}(\mathbf{A}\mathbf{A}^\top) = [\|\mathbf{a}_1\|_2^2, \|\mathbf{a}_2\|_2^2, \dots]^\top, \quad (31)$$

where $\mathbf{A}^\top = [\mathbf{a}_1, \mathbf{a}_2, \dots]$.

The second relation is for $\mathbf{A}\mathbf{x} = \mathbf{k} * \mathbf{x}$ (neglecting the boundary condition), we have

$$[\|\mathbf{a}_1\|_2^2, \|\mathbf{a}_2\|_2^2, \dots]^\top = \text{vec}(\|\mathbf{k}\|_2^2, \|\mathbf{k}\|_2^2 \dots) = \text{vec}(\mathbf{k}^{\cdot 2} * \mathbf{1}). \quad (32)$$

From (31) and (32), we have

$$\text{diag}(\mathbf{A}\mathbf{A}^\top) = \text{vec}(\mathbf{k}^{\cdot 2} * \mathbf{1}), \quad (33)$$

where \mathbf{A} is the convolution matrix corresponding to the PSF kernel \mathbf{k} (i.e., $\mathbf{A}\mathbf{x} = \mathbf{k} * \mathbf{x}$). A straight forward generalization leads to the following relation:

$$\text{diag}(\mathbf{A}\text{diag}(\mathbf{w})\mathbf{A}^\top) = \text{vec}(\mathbf{k}^{\cdot 2} * \mathbf{w}), \quad (34)$$

and

$$\text{diag}(\mathbf{A}^\top \text{diag}(\mathbf{w})\mathbf{A}) = \text{vec}(\bar{\mathbf{k}}^{\cdot 2} * \mathbf{w}), \quad (35)$$

where $\bar{\mathbf{k}} = \text{fliplr}(\text{flipud}(\mathbf{k}))$.

From (23) and (35), we have

$$\text{diag}(\mathbf{W}_\gamma) = \tau \text{vec}(\bar{\mathbf{h}} * \mathbf{D}^\top \mathbf{D} \mathbf{h} * \mathbf{1}) + \sum_{l=1}^L \text{vec}(\bar{\mathbf{k}}^{\cdot 2} * (\gamma_l^{-1})). \quad (36)$$

Recall the relationship between Σ and \mathbf{W}_γ as shown in (23), we have

$$\text{diag}(\Sigma_x) = \text{diag}(\tilde{\Sigma}_x) \approx 1./\text{diag}(\mathbf{W}_\gamma). \quad (37)$$

With this covariance approximation, we can now detail the computation of $\{\gamma_l\}_{l=1}^L$ as follows:

$$\begin{aligned} \gamma_l &= \text{diag}\langle \mathbf{u}_l \mathbf{u}_l^\top \rangle \\ &= (\boldsymbol{\mu}_{\mathbf{u}_l})^2 + \text{diag}(\Sigma_{\mathbf{u}_l}) \\ &= (\mathbf{K}_l \boldsymbol{\mu}_x)^2 + \text{diag}\{\mathbf{K}_l(\Sigma_x) \mathbf{K}_l^T\} \\ &\approx (\mathbf{K}_l \boldsymbol{\mu}_x)^2 + \text{diag}\{\mathbf{K}_l(\tilde{\Sigma}_x) \mathbf{K}_l^T\} \end{aligned} \quad (38)$$

where $\mathbf{x}^{\cdot 2}$ denotes the element-wise square of the vector \mathbf{x} . Plugging (37) into (38), we have the update equation for the latent variables $\{\gamma_l\}_{l=1}^L$:

$$\begin{aligned}\gamma_l &= \text{diag}\langle \mathbf{z}_l \mathbf{z}_l^T \rangle \\ &\approx (\mathbf{K}_l \boldsymbol{\mu}_x)^{\cdot 2} + \text{diag} \left\{ \mathbf{K}_l (\tilde{\Sigma}_x) \mathbf{K}_l^T \right\} \\ &= (\mathbf{K}_l \boldsymbol{\mu}_x)^{\cdot 2} + \text{vec}(\mathbf{k}_l^{\cdot 2} * \mathbf{v}),\end{aligned}\tag{39}$$

where $\mathbf{v} \triangleq \text{diag}(\Sigma_x)$. Similarly, from (29) and (37), the updating equation for w reduces to:

$$w = \sum_{i=1}^n [\text{vec}(\bar{\mathbf{h}} * \mathbf{D}^\top \mathbf{D} \mathbf{h} * \mathbf{v})]_i.\tag{40}$$

It is worthwhile to point out that the approximate computation of the covariance is only used for updating the parameters. The computation of the covariance is also involved in the estimation of the HR image \mathbf{x} in (22) but the covariance matrix there is not approximately calculated. As (22) can be solved efficiently with Conjugate Gradient (CG) method, it is unnecessary to construct \mathbf{K} and \mathbf{W}_z explicitly, as only the ability to evaluate a vector with respect to them is required in the CG algorithm, thus avoiding the intractability of explicitly constructing the covariance matrix or resorting to approximation to it.

Note that in (28), the prior knowledge on the noise level is incorporated into the final noise estimation via the shape and scale parameter a^o and b^o of the hyperprior. If we have no knowledge available on the noise level, we can simply use a uninformative prior by setting $a^o = 1$ and $b^o = 0$. The overall procedure of the proposed method is summarized in Algorithm 1. Note that when the downsampling operator \mathbf{D} is an identity matrix, our model reduces naturally to a deblurring algorithm.

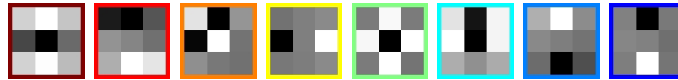


Fig. 2. Illustration of the 8 learned filters from natural images [45]. These filters are used in the FoE model with an infinite GSM as its expert functions. Note that the shape of this function is not determined beforehand but adaptively learned alongside the estimation process.

C. Implementation Details

In this subsection, we make some comments on the implementation details. The low-passing filter corresponding to \mathbf{H} is modeled as a Gaussian filter with standard deviation of 2 when the zooming factor is 3. The size of the Gaussian filter is set as 7. The filtering process is implemented via FFT in the

Algorithm 1: (Empirical Bayesian Super-Resolution.).

- 1: **Input:** LR observation \mathbf{y} , zooming factor r , number of iterations T .
 - 2: **Initialize:** set up \mathbf{D} and \mathbf{H} according to r and set the initial HR estimation as $\tilde{\mathbf{x}}_0 = \mathbf{H}^\top \mathbf{D}^\top \mathbf{y}$.
The γ_l 's are initialized as $\gamma_l = 10^{-5} \mathbf{1}$.
 - 3: **For** $i = 1, 2, \dots, T$, do
 - update the HR image \mathbf{x} via (22);
 - update the covariance approximation via (36) and (37);
 - update γ_l via (39);
 - update τ via (28) and (29).
 - 4: **End**
 - 5: **Output:** SR estimation \mathbf{x} .
-

frequency domain. The down sampling operator \mathbf{D} is implemented via subsampling the image for every r rows and columns for the zooming factor of r . Eight experts (thus $L = 8$) are multiplied together to form the energy potential $f(\cdot)$, which is a PoE model. Each expert is associated with one of the eight different filters as shown in Figure 2. These filters are used in the FoE model with a infinite GSM as its expert functions. Note that the shape of this function is not determined beforehand but adaptively learned alongside the estimation process. In our implementation, we use filters $\{\mathbf{k}_l\}_{l=1}^L$ with size of 3×3 . The filters $\{\mathbf{k}_l\}_{l=1}^L$ for each GSM are learned from the training images using the Contrastive Divergence algorithm [30], which are the same as those used in our previous work [1]. Specifically, we use the implementation given by [45] which is trained on the Berkeley Segmentation database [46]. Note that the training dataset does not include any test images used for comparison in the experimental section.

V. EXPERIMENTS AND RESULTS

In this section, we conduct several experiments to evaluate the effectiveness of the proposed method, both qualitatively and quantitatively. We first give an illustrative example to demonstrate some basic properties of the proposed algorithm. Then we compare the SR estimation results with several classical as well as *state-of-the-art* SR methods, on several standard test images as well as real-world color images. Finally, the robustness of the proposed SR method with respect to noise is also investigated. Peak Signal to Noise Ratio (PSNR) is adopted as an evaluation metric which is defined as:

$$\text{PSNR} = 10 \log_{10} \frac{255^2}{\frac{1}{m} \sum_i (\hat{\mathbf{x}} - \mathbf{x})_i^2},$$

TABLE I
SUPER RESOLUTION ($\times 3$) RESULT COMPARISON OF ESTIMATION QUALITY.

Methods		NN	BI	Fast [19]	ScSR [6]	NBSR [1]	Proposed
house	PSNR	24.90	25.62	28.72	30.80	32.06	32.85
	SSIM	0.763	0.788	0.837	0.835	0.897	0.897
peppers	PSNR	21.59	22.18	24.38	25.52	25.88	26.88
	SSIM	0.727	0.775	0.841	0.863	0.910	0.918
cameraman	PSNR	21.71	22.12	24.64	25.79	26.42	26.65
	SSIM	0.695	0.714	0.774	0.810	0.847	0.846
barbara	PSNR	23.07	23.43	24.58	25.24	25.62	25.60
	SSIM	0.618	0.649	0.683	0.732	0.753	0.749
lena	PSNR	25.91	26.62	29.59	31.89	33.25	33.40
	SSIM	0.769	0.802	0.833	0.888	0.911	0.909
boat	PSNR	24.14	24.63	26.70	28.25	29.47	29.34
	SSIM	0.659	0.687	0.729	0.815	0.844	0.836
hill	PSNR	25.88	26.41	27.72	30.15	30.44	30.84
	SSIM	0.654	0.686	0.714	0.811	0.831	0.823
couple	PSNR	24.16	24.60	26.34	27.69	28.53	28.64
	SSIM	0.621	0.649	0.698	0.788	0.811	0.811

where \hat{x} denotes the estimated SR image and x denotes the original HR image. m is the total number of pixels in x . The Structural SIMilarity index (SSIM) is used as another objective measure, which aligns better with the human perception than PSNR [47].

A. Super Resolution Experiment

To verify the effectiveness of the proposed method, we compare the SR performance of the proposed method with several algorithms: Nearest Neighbor interpolation (NN), bicubic interpolation (BI), fast SR method [19], ScSR method [6],³ Variational Bayesian based SR (VBSR) method [16]⁴, as well as the Natural image prior based Bayesian SR method using sampling (NBSR) [1]. In this experiment, we use the standard test images in the noise-free setting for evaluation and use PSNR and SSIM as the objective measures. The zooming factor in this experiment is $r = 3$. The SR results in terms of PSNR

³For the ScSR method, the dictionary size is 1024 and the regularization parameter is set as $\lambda = 0.1$ which is the same setting as in [6].

⁴codes are available at <http://decsai.ugr.es/pi/superresolution/software.html>



Fig. 3. Single image super resolution results for ‘peppers’, ‘cameraman’ and ‘lena’ images ($\times 3$) with different algorithms: Nearest Neighbor Interpolation, sparse representation based SR method [6], the NBSR method [1] and the proposed EBSR method. The results evaluated in terms of PSNR and SSIM are reported in Table I.

and SSIM are summarized in Table I. As can be seen from Table I, the proposed method can generate SR images with better quality in terms of PSNR and SSIM than the fast SR method [19] and the ScSR method [6] on all the testing images. Also, the proposed method performs comparable or better than the NBSR [1] method. This clearly demonstrates the effectiveness of the proposed method. To further compare the SR results visually, we show some of the SR results from different algorithms in Figure 3. As can be seen from Figure 3, the fast SR method tends to removing the fine details from the image and the SR results exhibit cartoon-like artifacts. Furthermore, it is noticed from Figure 3 that the SR results from the ScSR method have some artifacts along the edge structures (see the zoomed patches for ‘peppers’ and ‘cameraman’ images in Figure 3). This may result from the inconsistency in recovering the HR patches from the neighboring LR patches. On the other hand, the SR image from the proposed method does not suffer from these artifacts, due to the usage of the global statistical model as well as

TABLE II
SUPER RESOLUTION ($\times 4$) QUALITY COMPARISON ($\sigma = 1$).

Methods		NN	BI	Fast [19]	KRR [11]	ScSR [6]	VBSR [16]	NBSR [1]	Proposed
house	PSNR	22.24	23.14	26.79	27.60	26.86	27.00	28.20	28.81
	SSIM	0.627	0.671	0.771	0.785	0.742	0.767	0.767	0.802
peppers	PSNR	21.47	22.20	25.36	25.95	25.61	25.36	26.90	27.51
	SSIM	0.647	0.711	0.786	0.821	0.764	0.807	0.837	0.841
cameraman	PSNR	18.35	18.91	21.06	21.35	21.17	21.50	21.72	22.43
	SSIM	0.695	0.714	0.742	0.758	0.695	0.725	0.760	0.794
barbara	PSNR	21.91	21.50	24.12	24.75	24.62	24.12	24.94	25.15
	SSIM	0.515	0.582	0.640	0.696	0.654	0.699	0.703	0.701
lena	PSNR	21.68	22.41	26.66	26.88	27.17	27.62	28.32	29.25
	SSIM	0.624	0.678	0.766	0.807	0.760	0.814	0.825	0.829
boat	PSNR	19.52	20.17	22.47	23.27	22.94	23.81	23.57	24.36
	SSIM	0.659	0.687	0.614	0.652	0.598	0.657	0.640	0.686
hill	PSNR	24.56	25.17	26.94	27.75	24.92	27.10	27.94	28.44
	SSIM	0.540	0.581	0.623	0.666	0.669	0.675	0.680	0.682
couple	PSNR	21.98	22.41	24.08	24.79	25.17	25.59	25.11	25.53
	SSIM	0.524	0.564	0.655	0.697	0.641	0.707	0.663	0.719

the MMSE estimation criteria for the latent HR images. This clearly demonstrates the benefit of using a global statistical model as a prior based on natural image statistics, combined with empirical Bayesian estimation scheme.

To further verify the effectiveness of the proposed method, we compare its SR performance with more SR algorithms, including the Kernel Ridge Regression SR (KRR) method [11], the Variational Bayesian SR (VBSR) method [16] and NBSR [1], with zooming factor of 4.⁵ The KRR method achieves SR via patch-based sparse kernel regression for capturing the mapping between low resolution and high resolution patches. The VBSR method uses a simple TV prior model and estimates the HR image with posterior mean via variational Bayesian approximation. As in the current implementation of the VBSR method, the matrices corresponding to the low-pass filtering as well as the down-sampling operator are constructed explicitly, thus it can only handle small images. Therefore, we crop image parts of size 192×192 from the standard test images and summarize the results in Table II. As can be seen from Table II, the KRR

⁵As both the KRR and VBSR methods can only handle SR tasks with even zooming factor, they are not compared in Table I.

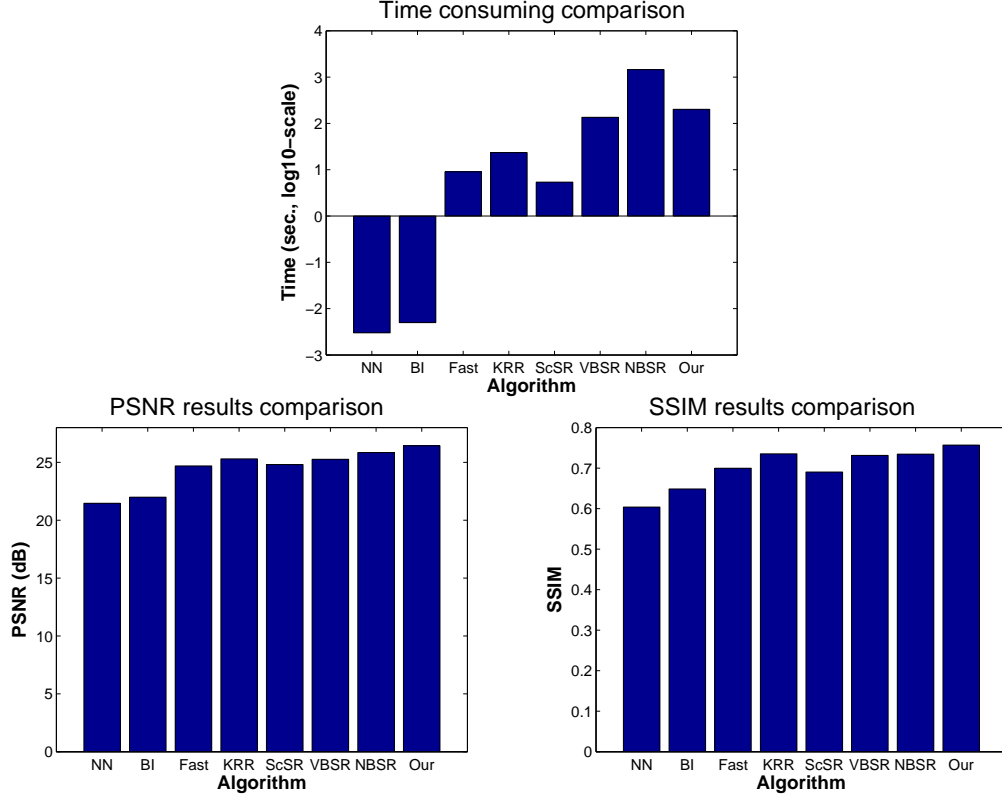


Fig. 4. Computational time consuming comparison and the corresponding SR performs comparison in terms of PSNR and SSIM.

method performs better than the fast SR method in general, and the VBSR method overall performs similar to the ScSR method and the KRR method. Finally, the proposed method outperforms the VBSR method and all the other methods constantly in terms of both PSNR and SSIM, which demonstrates the superiority of the proposed method. The average computational time for different algorithms as well as their corresponding SR performance in terms of PSNR and SSIM are shown in Figure 4.⁶ The typical number of iteration is 10 for the fast SR method, 30 for VBSR, 100 for NBSR and 10 for the proposed method. The other algorithms in Table II are ‘one-shot’ methods that require no iteration. As can be observed from Figure 4 that proposed method outperforms the state-of-art NBSR [1] method in terms of HR estimation quality while can reduce the computational time by around an order of magnitude. Besides, the proposed method has similar computational cost with the VBSR [16] while the proposed method performs much better.

⁶The computational time is measured on a Laptop with Intel Core2 Duo P8600 (2.4GHz) CPU and 1GB RAM.

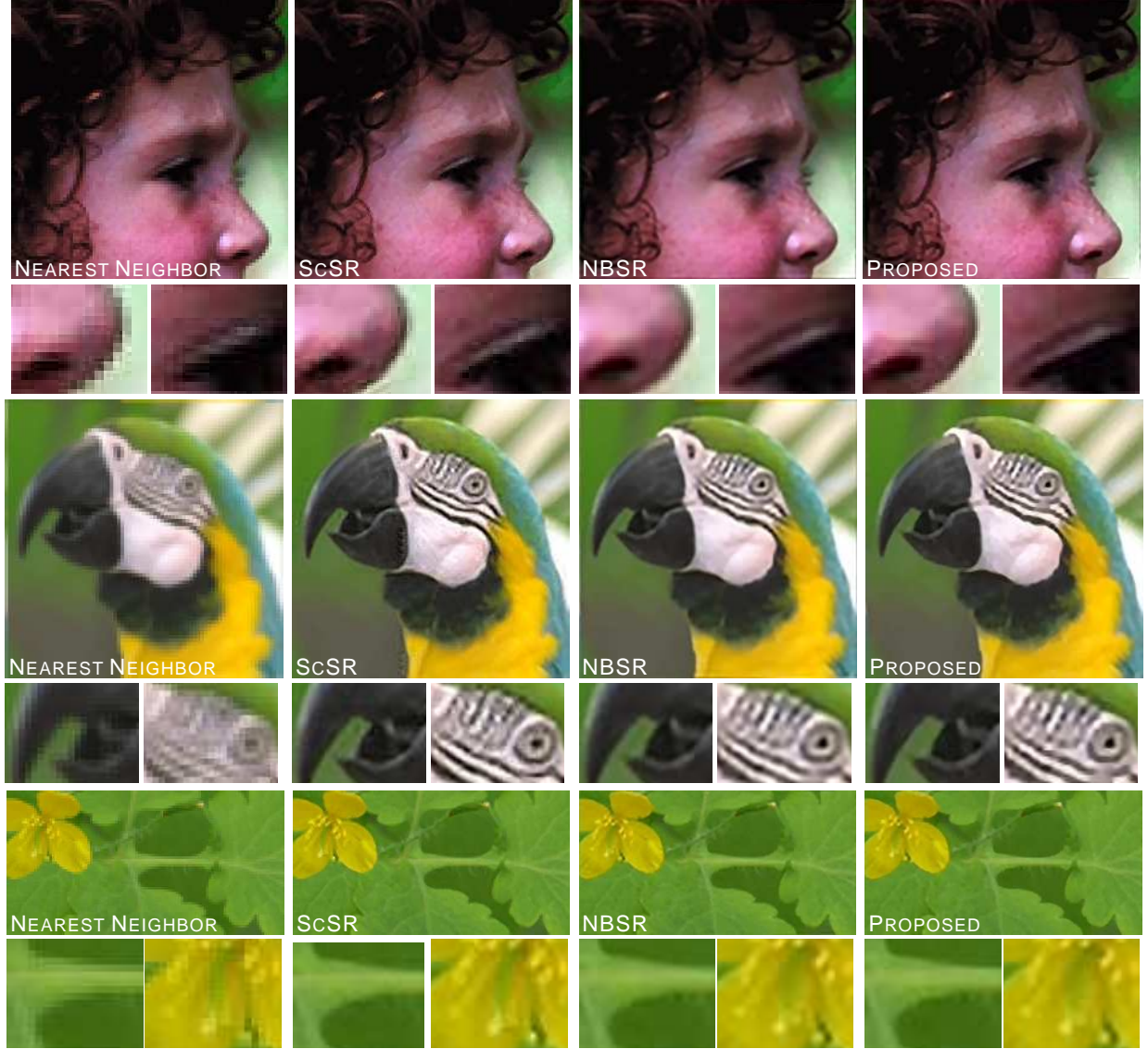


Fig. 5. Single image super resolution ($\times 3$) results for ‘girl’, ‘parrot’ and ‘flower’ images ($\times 3$). PSNR and SSIM results for the image ‘flower’ are given in brackets. Left to right: Nearest Neighbor Interpolation (32.95, 0.849), Fast super resolution method [19] (35.81, 0.890), sparse representation based SR method [6] (37.60, 0.923), the NBSR method [1] (38.07, **0.927**) and the proposed EBSR method (**38.30, 0.927**).

We further evaluate the proposed SR method on the task of color image SR. For color image SR, we only perform SR for the intensity channel in the YCbCr color space and use bicubic interpolation for the other two channels. The reason is that human eyes are more sensitive to the intensity changes of the image than to the changes in other channels such as saturation and hue. This scheme has been practised already in the previous SR works [6]. The SR results on color images with the zooming factor of 3 are

TABLE III
NOISY IMAGE SUPER RESOLUTION ($\times 3$) RESULT COMPARISON OF ESTIMATION QUALITY.

σ	Methods		NN	BI	Fast [19]	ScSR [6]	NBSR [1]	Proposed
0	house	PSNR	24.90	25.62	28.72	30.80	32.06	32.85
		SSIM	0.763	0.788	0.837	0.863	0.897	0.897
	peppers	PSNR	21.59	22.18	24.38	25.52	25.88	26.88
		SSIM	0.727	0.775	0.841	0.863	0.910	0.918
	cameraman	PSNR	21.71	22.12	24.64	25.79	26.42	26.65
		SSIM	0.695	0.714	0.774	0.810	0.847	0.846
1	house	PSNR	24.88	25.60	28.88	30.67	31.24	32.13
		SSIM	0.757	0.784	0.844	0.854	0.876	0.879
	peppers	PSNR	21.58	22.17	24.53	25.49	25.65	26.37
		SSIM	0.723	0.772	0.851	0.858	0.886	0.894
	cameraman	PSNR	21.70	22.15	24.73	25.75	26.09	26.32
		SSIM	0.689	0.710	0.778	0.801	0.823	0.828
2	house	PSNR	24.81	25.55	28.76	30.47	30.47	31.33
		SSIM	0.738	0.773	0.840	0.834	0.858	0.864
	peppers	PSNR	21.55	22.15	24.49	25.36	25.23	25.84
		SSIM	0.710	0.764	0.847	0.844	0.860	0.871
	cameraman	PSNR	21.66	22.13	24.70	25.63	25.67	25.92
		SSIM	0.671	0.700	0.775	0.779	0.802	0.809
4	house	PSNR	24.57	25.36	28.41	29.38	29.33	30.07
		SSIM	0.678	0.734	0.823	0.798	0.832	0.842
	peppers	PSNR	21.44	22.06	24.29	24.65	24.54	25.01
		SSIM	0.667	0.736	0.830	0.803	0.823	0.834
	cameraman	PSNR	21.54	22.03	24.52	24.88	24.94	25.18
		SSIM	0.612	0.662	0.757	0.734	0.770	0.778

shown in Figure 5. The PSNR and SSIM results accompanying these images are also reported. As can be seen from Figure 5, the proposed method can generate SR images visually comparable to the sparse representation based method, which are much better than the fast SR method. Detailed inspections also reveal that the SR results from the proposed method have less artifacts than those from the ScSR method (see the zoomed patches for the *nose* and *eye* parts of the ‘girl’ image in Figure 5), and appear sharper than the results from the NBSR method. Furthermore, the quantitative evaluation also indicates that the SR results from the proposed method have better quality in terms of PSNR and SSIM.

More color image SR results are shown in Figure 6 and Figure 7, with the zooming factor of 4.

We compare the SR result from our proposed method with the results from several SR methods in the literature: the results from Freeman’s example-based SR method [3], Kim’s kernel regression method [11], Glasner’s method [4], Shan *et al.*’s fast SR method [19], Yang *et al.*’s sparse representation based SR method [6] as well as the Variational Bayesian based SR method [16]. From Figure 6, we can see that the fast SR method again over-smooths the HR image, and the ScSR method generates HR images with artifacts along the edges. The HR image from the VBSR method suffers from ringing artifacts. The proposed method, on the other hand, generates HR image with less artifacts and with desirable quality both visually and in terms of PSNR and SSIM, and with comparable (slightly better) quality than the results of NBSR. More SR algorithms are compared in Figure 7. It can be observed from Figure 7 that the SR results from Freeman’s example-based SR method and Glasner’s method although have visual resolution improvement, they actually hallucinating some high resolution information that is not fidelity to the original image, thus hindering the improvement in terms of PSNR and SSIM. Fattal’s method and Shan’s method appear to suffer from cartoon-like artifacts. The proposed method is visually much better than results via NN; also, it has less artifacts compared to the result from example-based SR method [3], and is sharper than result from Kim’s kernel regression method [11] and the NBSR method [1], and achieves the best performance in terms of PSNR and SSIM. The results on both gray and color image SR all verified the effectiveness of the proposed method.

B. Noisy Image Super Resolution

In real-world SR tasks, the observed LR images are rarely noise-free, but are often contaminated by noise. Therefore, it is important for the SR algorithm to be robust to the noise contained in the LR observations. In this subsection, we evaluate the effectiveness of the proposed approach in the case of noisy SR situation, *i.e.*, the given input is both of low in resolution and is contaminated by noise. The SR estimation results under the noise levels $\sigma \in \{0, 1, 2, 4\}$ are reported in Table III. The bar plots for the evaluated algorithms are shown in Figure 8. The PSNR and SSIM performance shown in Figure 8 is the average performance over the test images. As can be seen from Table III and Figure 8, the proposed method performs better than the other methods when the noise level is low; as the noise level increases, the performances in terms of PSNR and SSIM measures of all the SR algorithms decreases. In this case, the performance of the proposed method is still better than the other algorithms in general in terms of SSIM and is comparable to ScSR in terms of PSNR, indicating its robustness to the noise in the LR images and its applicability to general SR scenarios. It is worthwhile to point out that for the ScSR method, it is given the ground truth noise level, and its regularization weight has to be adjusted

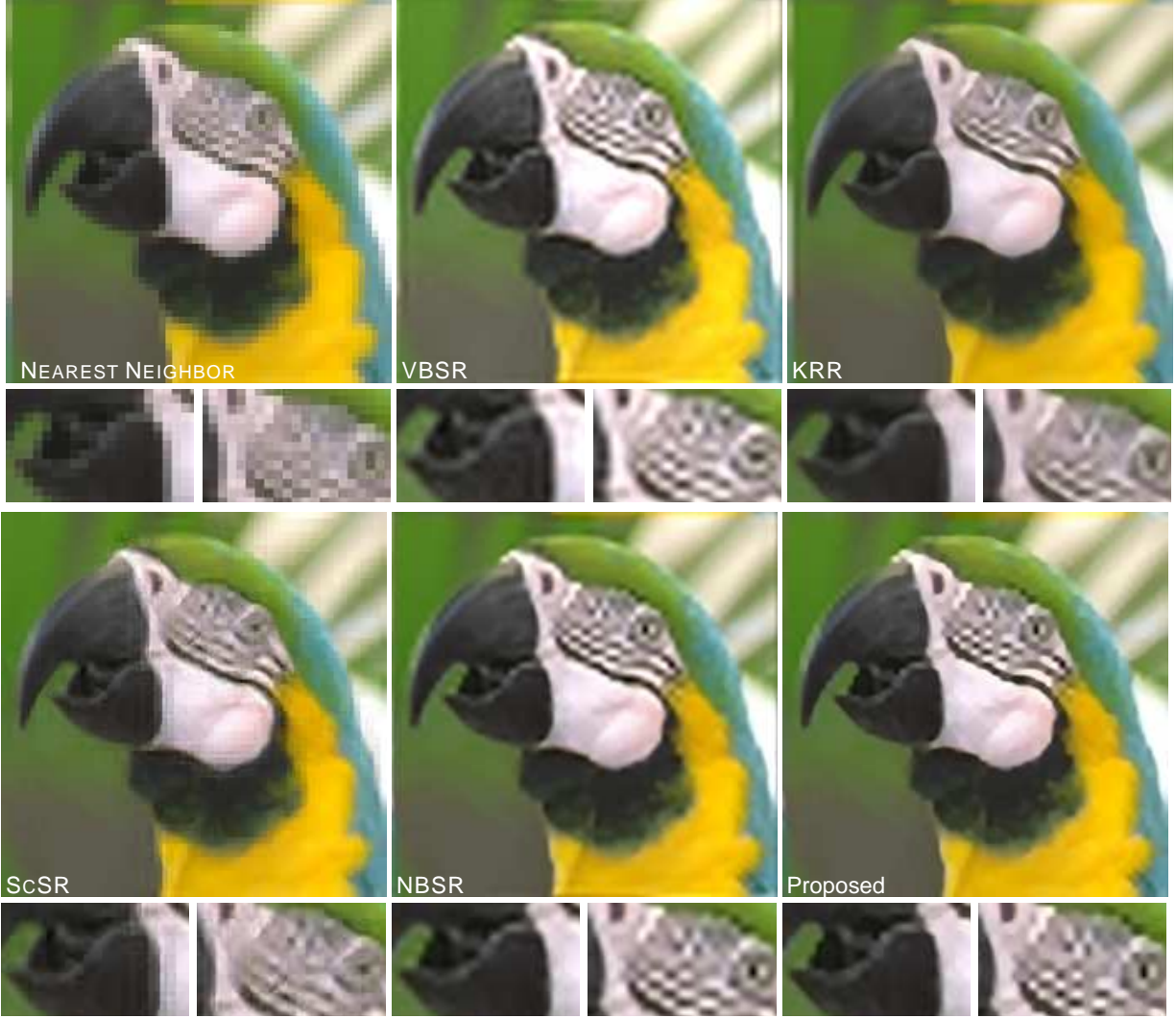


Fig. 6. Single image super resolution results with different algorithms ($\times 4$). The results are generated with: Nearest Neighbor interpolation (22.64, 0.699), Shan *et al.*'s Fast SR method [19] (24.09, 0.775), VBSR [16] (23.33, 0.823), Kim *et al.*'s kernel regression based method (24.22, 0.818) [11], Yang *et al.*'s ScSR method (25.47, 0.798) [6], Zhang *et al.*'s NBSR method [1] (26.22, 0.847) and the proposed EBSR method (**26.47, 0.849**).

according to the noise level.⁷ The proposed method, on the other hand, does not require the noise level to be known, but can infer it from the noise LR image itself, while generating results comparable or better than those from ScSR, demonstrating its effectiveness for real-world SR tasks.

⁷In our experiments, we adjust the regularization weight the same way as suggested in [6].



Fig. 7. Single image super resolution results with different algorithms ($\times 4$). The results are generated with: Nearest Neighbor interpolation (26.05, 0.741), Bicubic interpolation (26.92, 0.785), Freeman *et al.*'s example based SR method (25.38, 0.729), Kim *et al.*'s kernel regression based method (31.22, 0.863) [11], Fattal *et al.*'s edge statistics-based SR method (25.55, 0.751) [12], Glasner *et al.*'s method (25.28, 0.741) [4], Shan *et al.*'s method (24.43, 0.727) [19], NBSR [1] (31.80, 0.8846) and the proposed EBSR method (**33.00, 0.8854**).

VI. CONCLUSION

An effective and efficient Bayesian image SR algorithm with natural image prior is proposed in this paper. The proposed method exploits the natural image statistics for image SR with a flexible high-order MRF model. Specifically, an FoE model with products of GSM potentials is used for learning the prior model from natural images, which is further incorporated into a fully Bayesian framework for image SR. Moreover, the GSM is adapted to the content of the HR image alongside with the estimation process of the HR image. To estimate the latent HR image, we use the empirical Bayesian approach, which can generate high quality estimation while avoiding the time-consuming sampling. Experimental results under different settings compared with several *state-of-the-art* SR algorithms in the literature verify

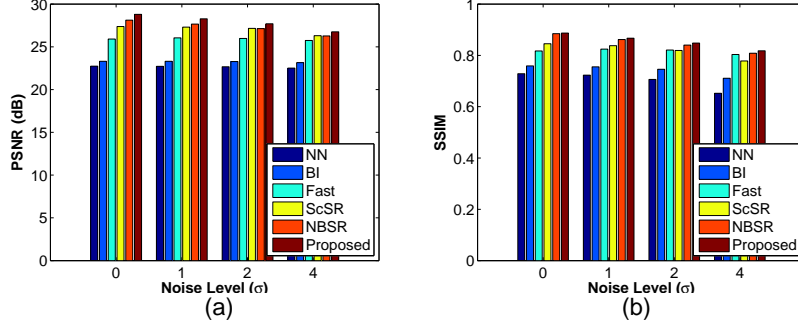


Fig. 8. Image SR results. (a) PSNR bar plot and (b) SSIM bar plot with respect to increasing noise level ($\sigma \in \{0, 1, 2, 4\}$).

the effectiveness of the proposed method. Specifically, compared with our previous sampling based SR approach [1], the method proposed in this paper can generate SR results with comparable or better quality than our previous sampling based SR approach [1], while with far less computational cost. For future work, we would like to extend our current approach to other related inverse problems, such as multi-frame super resolution. The extension of the current work to blind deblurring is our current ongoing work, which is encouraging based on our now obtained results.

REFERENCES

- [1] H. Zhang, Y. Zhang, H. Li, and T. S. Huang, "Generative bayesian image super resolution with natural image prior," *IEEE Trans. on Image Processing*, 2012.
- [2] R.Y. Tsai and T. Huang, "Multiframe image restoration and registration," pp. 317–339, 1984.
- [3] W.T. Freeman, T.R. Jones, and E.C. Pasztor, "Example based super-resolution," in *IEEE computer graphics and applications*, 2002.
- [4] D. Glasner, S. Bagon, and M. Irani, "Super-resolution from a single image," in *IEEE International Conference on Computer Vision (ICCV)*, 2009, pp. 349–356.
- [5] H. Zhang, J. Yang, Y. Zhang, and T.S. Huang, "Non-local kernel regression for image and video restoration," in *European Conference on Computer Vision (ECCV)*, 2010, pp. 566–579.
- [6] J. Yang, J. Wright, T. Huang, and Y. Ma, "Image super-resolution via sparse representation," *IEEE Transactions on Image Processing*, vol. 19, no. 11, pp. 2861–2873, 2010.
- [7] S. Baker and T. Kanade, "Limits on super-resolution and how to break them," *IEEE Transactions on Pattern Analysis and Machine Intelligence*, vol. 24, no. 9, pp. 1167–1183, 2002.
- [8] X. Li and M.T. Orchard, "New edge-directed interpolation," *IEEE Transactions on Image Processing*, vol. 10, no. 10, pp. 1521–1527, 2001.
- [9] H. Takeda, S. Farsiu, and P. Milanfar, "Kernel regression for image processing and reconstruction," *IEEE Transactions on Image Processing*, vol. 16, no. 2, pp. 349–366, 2007.
- [10] M.F. Tappen, B.C. Russell, and W.T. Freeman, "Exploiting the sparse derivative prior for super-resolution and image demosaicing," in *In IEEE Workshop on Statistical and Computational Theories of Vision*, 2003.

- [11] K.I. Kim and Y. Kwon, "Single-image super-resolution using sparse regression and natural image prior," *IEEE Transactions on Pattern Analysis and Machine Intelligence*, vol. 32, no. 6, pp. 1127–1133, 2010.
- [12] R. Fattal, "Image upsampling via imposed edge statistics," *ACM Transactions on Graphics (SIGGRAPH)*, vol. 26, no. 3, pp. 95–102, 2007.
- [13] S. Dai, M. Han, W. Xu, Y. Wu, Y. Gong, and A.K. Katsaggelos, "Softcuts: A soft edge smoothness prior for color image super-resolution," *IEEE Transactions on Image Processing*, vol. 18, no. 5, pp. 969–981, 2009.
- [14] J. Sun, J. Sun, Z. Xu, and H.Y. Shum, "Image super-resolution using gradient profile prior," in *IEEE Conference on Computer Vision and Pattern Recognition (CVPR)*, 2008, pp. 1–8.
- [15] A. Mohebi and P. Fieguth, "Posterior sampling of scientific images," in *International Conference on Image Analysis and Recognition*, 2006.
- [16] S.D. Babacan, R. Molina, and A. K. Katsaggelos, "Variational bayesian super resolution," *IEEE Transactions on Image Processing*, vol. 20, no. 4, pp. 984–999, 2011.
- [17] T. Katsuki, A. Torii, and M. Inoue, "Posterior mean super-resolution with a causal gaussian markov random field prior," *CoRR*, vol. abs/1109.5453, 2011.
- [18] M.E. Tipping and C.M. Bishop, "Bayesian image super-resolution," in *Advances in Neural Information Processing Systems (NIPS)*, 2005, pp. 1303–1310.
- [19] Q. Shan, Z. Li, J. Jia, and C.K. Tang, "Fast image/video upsampling," in *ACM Transactions on Graphics (SIGGRAPH ASIA)*, 2008, vol. 27.
- [20] Y.W. Tai, S. Liu, M.S. Brown, and S. Lin, "Super resolution using edge prior and single image detail synthesis," in *IEEE Conference on Computer Vision and Pattern Recognition (CVPR)*, 2010, pp. 2400–2407.
- [21] L.C. Pickup, S.J. Roberts, and A. Zisserman, "A sampled texture prior for image super-resolution," in *Advances in Neural Information Processing Systems (NIPS)*, 2003, pp. 1587–1594.
- [22] Y. HaCohen, R. Fattal, and D. Lischinski, "Image upsampling via texture hallucination," in *IEEE International Conference on Computational Photography (ICCP)*, 2010, pp. 1–8.
- [23] D.J. Field, "Relations between the statistics of natural images and the response profiles of cortical cells," *Journal of the Optical Society of America*, vol. 4, no. 12, pp. 2379–2394, 1987.
- [24] M. Zontak and M. Irani, "Internal statistics of a single natural image," in *IEEE Conference on Computer Vision and Pattern Recognition (CVPR)*, 2011, pp. 977–984.
- [25] Y. Weiss and W.T. Freeman, "What makes a good model of natural images," in *IEEE Conference on Computer Vision and Pattern Recognition (CVPR)*, 2007, pp. 1–8.
- [26] S. Roth and M.J. Black, "Fields of experts," *International Journal of Computer Vision*, vol. 82, no. 2, pp. 205–229, 2009.
- [27] R. Fergus, B. Singh, A. Hertzmann, S.T. Roweis, and W.T. Freeman, "Removing camera shake from a single photograph," *ACM Transactions on Graphics (SIGGRAPH)*, vol. 25, pp. 787–794, 2006.
- [28] S. Osher, M. Burger, and D. Goldfarb, "An iterative regularization method for total variation-based image restoration," *Multiscale Modeling & Simulation*, vol. 4, pp. 460–489, 2005.
- [29] D. Krishnan and R. Fergus, "Fast image deconvolution using hyper-laplacian priors," in *Advances in Neural Information Processing Systems (NIPS)*, 2009, pp. 1033–1041.
- [30] G.E. Hinton, "Training products of experts by minimizing contrastive divergence," *Neural Computation*, vol. 14, no. 8, pp. 1771–1800, 2002.

- [31] T.S. Cho, C.L. Zitnick, N. Joshi, S. B. Kang, R. Szeliski, and W.T. Freeman, "Image restoration by matching gradient distributions," *IEEE Transactions on Pattern Analysis and Machine Intelligence*, vol. 34, no. 4, pp. 683–694, 2012.
- [32] S. Lyu, "Interpretation and generalization of score matching," in *the Twenty-Fifth Conference Annual Conference on Uncertainty in Artificial Intelligence (UAI-09)*. 2009, pp. 359–366, AUAI Press.
- [33] J. Miskin and D.J.C. Mackay, "Ensemble learning for blind image separation and deconvolution," in *Advances in Independent Component Analysis*, M. Girolami, Ed. 2000, Springer-Verlag Scientific Publishers.
- [34] L. He, H. Chen, and L. Carin, "Tree-structured compressive sensing with variational bayesian analysis," *IEEE Signal Processing Letters*, vol. 17, no. 3, pp. 233–236, 2010.
- [35] M. E. Tipping, "Sparse bayesian learning and the relevance vector machine," *Journal of Machine Learning Research*, vol. 1, pp. 211–244, 2001.
- [36] D. P. Wipf, B. D. Rao, and S. S. Nagarajan, "Latent variable bayesian models for promoting sparsity," *IEEE Transactions on Information Theory*, vol. 57, no. 9, pp. 6236–6255, 2011.
- [37] D. P. Wipf and B. D. Rao, "Sparse bayesian learning for basis selection," *IEEE Transactions on Signal Processing*, pp. 2153–2164, 2004.
- [38] D. P. Wipf and B. D. Rao, "An empirical bayesian strategy for solving the simultaneous sparse approximation problem," *IEEE Transactions on Signal Processing*, vol. 55, no. 7-2, pp. 3704–3716, 2007.
- [39] D. P. Wipf, "Sparse estimation with structured dictionaries," in *NIPS*, 2011, pp. 2016–2024.
- [40] D. P. Wipf and S. S. Nagarajan, "Iterative reweighted ℓ_1 and ℓ_2 methods for finding sparse solutions," *IEEE J. Sel. Topics Signal Processing*, vol. 4, no. 2, pp. 317–329, 2010.
- [41] D. P. Wipf and Y. Wu, "Dual-space analysis of the sparse linear model," in *submitted to NIPS*, 2012.
- [42] M. W. Seeger, "Bayesian inference and optimal design for the sparse linear model," *Journal of Machine Learning Research*, vol. 9, pp. 759–813, 2008.
- [43] G. Papandreou and A. Yuille, "Efficient variational inference in large-scale bayesian compressed sensing," in *Proc. IEEE Workshop on Information Theory in Computer Vision and Pattern Recognition (in conjunction with ICCV)*, Barcelona, Spain, Nov. 2011, pp. 1332–1339.
- [44] Y. J. Ko and M. Seeger, "Large scale variational bayesian inference for structured scale mixture models," in *International Conference on Machine Learning (ICML)*, 2012.
- [45] U. Schmidt, Q. Gao, and S. Roth, "A generative perspective on MRFs in low-level vision," in *IEEE Conference on Computer Vision and Pattern Recognition (CVPR)*, 2010, pp. 1751–1758.
- [46] D. Martin, C. Fowlkes, D. Tal, and J. Malik, "A database of human segmented natural images and its application to evaluating segmentation algorithms and measuring ecological statistics," in *IEEE International Conference on Computer Vision (ICCV)*, 2001, vol. 2, pp. 416–423.
- [47] Z. Wang, A.C. Bovik, H.R. Sheikh, and E.P. Simoncelli, "Image quality assessment: From error visibility to structural similarity," *IEEE Transactions on Image Processing*, vol. 13, no. 4, pp. 600–612, 2004.